Calibrated Language Models Must Hallucinate

Year: 2023 | Citations: 125 | Authors: A. Kalai, S. Vempala

Abstract

Recent language models generate false but plausible-sounding text with surprising frequency. Such “hallucinations” are an obstacle to the usability of language-based AI systems and can harm people who rely upon their outputs. This work shows that there is an inherent statistical lower-bound on the rate that pretrained language models hallucinate certain types of facts, having nothing to do with the transformer LM architecture or data quality. For “arbitrary” facts whose veracity cannot be determined from the training data, we show that hallucinations must occur at a certain rate for language models that satisfy a statistical calibration condition appropriate for generative language models. Specifically, if the maximum probability of any fact is bounded, we show that the probability of generating a hallucination is close to the fraction of facts that occur exactly once in the training data (a “Good-Turing” estimate), even assuming ideal training data without errors. One conclusion is that models pretrained to be sufficiently good predictors (i.e., calibrated) may require post-training to mitigate hallucinations on the type of arbitrary facts that tend to appear once in the training set. However, our analysis also suggests that there is no statistical reason that pretraining will lead to hallucination on facts that tend to appear more than once in the training data (like references to publications such as articles and books, whose hallucinations have been particularly notable and problematic) or on systematic facts (like arithmetic calculations). Therefore, different architectures and learning algorithms may mitigate these latter types of hallucinations.